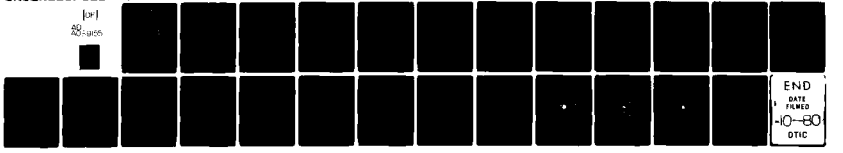
AD-A089 155 PRINCETON UNIV NJ DEPT OF STATISTICS F/G 12/1
DIRECTIONAL DATA ANALYSIS.(U)
JUL 80 G S WATSON N00014-79-C-0322
UNCLASSIFIED TR-170-SFR-2 NL



DIRECTIONAL DATA ANALYSIS

by

G.S. Watson
Princeton University

Technical Report No. 170, Series 2
Department of Statistics
Princeton University
July 1980

Research supported in part by a contract with the Office of Naval Research, No. N00014-79-C-0322, awarded to the Department of Statistics, Princeton University, Princeton, New Jersey.

A B S T R A C T

This article, written for the *Encyclopaedia of Statistical Sciences*, gives a brief introduction with many references to directional data analysis.

Key Words: Direction, Orientation, Fisher Distribution, Von Mises Distribution, Tests of Uniformity.

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report No. 170	2. GOVT ACCESSION NO.: AD-A089155
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: Directional Data Analysis	5. TYPE OF REPORT & PERIOD COVERED: Technical
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Geoffrey S. Watson	8. CONTRACT OR GRANT NUMBER(s): N00014-79-C-0322
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Princeton University, Princeton, NJ 08544	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research (Code 436), Arlington, VA 22217	12. REPORT DATE: July 1980
	13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)*
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Direction, Orientation, Fisher Distribution, Von Mises Distribution, Tests of Uniformity

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

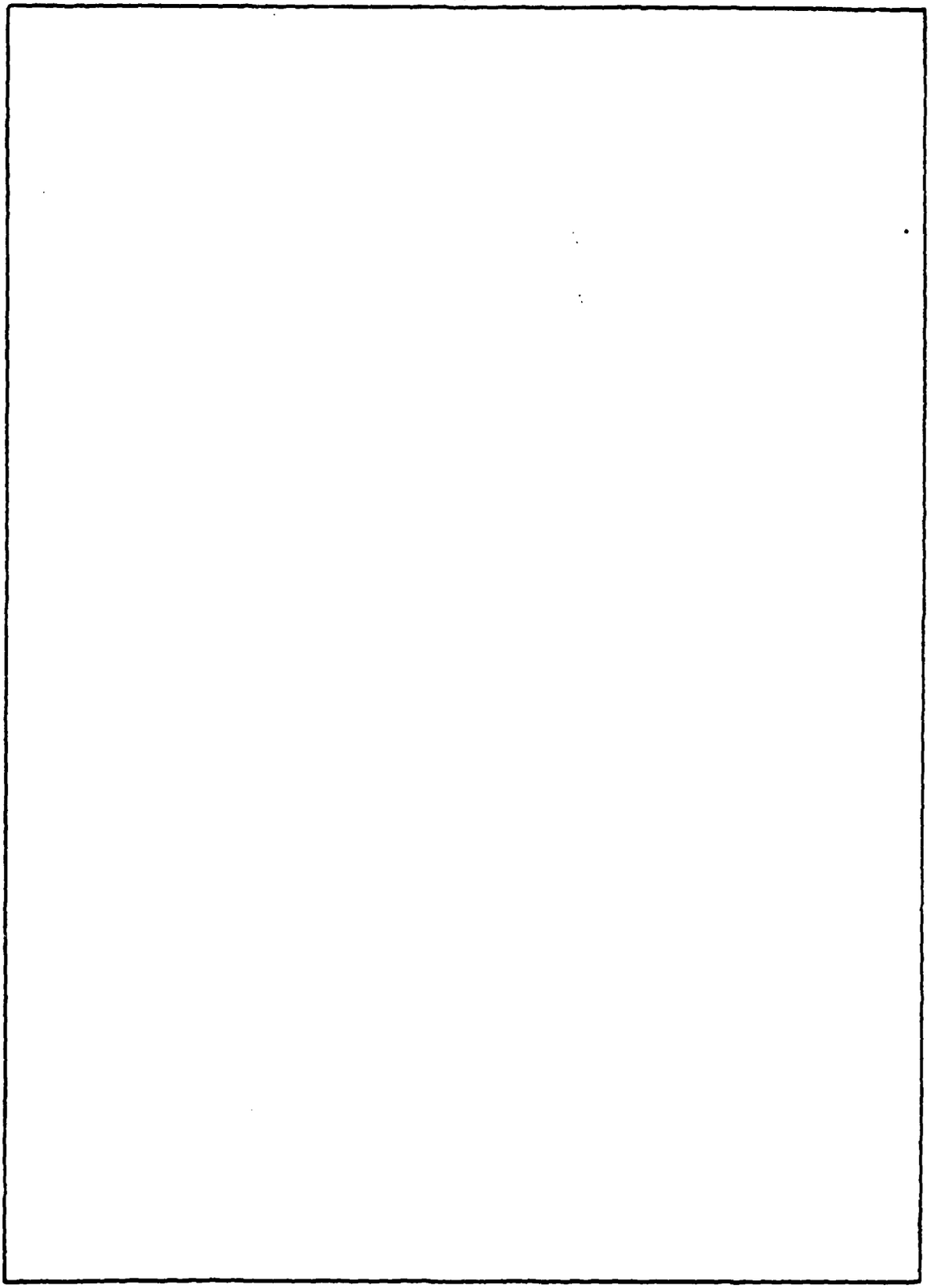
This article, written for the Enchclopaedia of Statistical Sciences, gives a brief introduction with many references to directional data analysis.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

4/06/15



1. INTRODUCTION

The current methods of analyzing directional data were motivated by measurements of (i) the direction of the long axis of pebbles (Krumbein, 1939), (ii) the direction (not strength) of magnetization of rocks (Fisher, 1953), (iii) the vanishing bearings of homing pigeons. In case (i) we have axial, not directional, data, and the axes could be oriented in two or three dimensions. The data could be displayed by marking the two points where the axis cuts a unit circle or sphere. In case (ii), each measurement can be thought of as a unit vector or as a point on a sphere of unit radius. In case (iii), each measurement can be thought of as an angle, a point on a circle of unit radius or a unit vector in the plane. In each of these cases, the sample of axes or directions has a fairly symmetric cluster about some "mean" direction, so that some scalar might be sought to describe the "dispersion" of the data.

Thus we may seek for directions, analogues of the mean and variance of data on the real line -- and even of the normal distribution. The distribution used is known by the names von Mises on the circle, and Fisher on the sphere and higher dimensions. This density is proportional to $\exp K \cos \theta$

where θ is the angle between the population mean direction and the direction of an observation. $K \geq 0$ is an accuracy parameter. For some axial data, the density

proportional to exp K $\cos^2 \theta$ is helpful. The scatter of data on the sphere in some applications suggests more general densities of the form (Beran (1974))

$$\exp \sum s_r S_r(\theta,\phi)$$

where $S_r(\theta,\phi)$ is a surface harmonic of degree r in spherical polar form. The special case r=1 is the Fisher distribution; that with S_2 only is the Bingham (1974) distribution which may be written

$$\exp \{K_1 (\underline{r}'\underline{\mu}_1)^2 + K_2 (\underline{r}'\underline{\mu}_2)^2 + K_3 (\underline{r}'\underline{\mu}_3)^2\}$$

where the terms $\underline{r}'\underline{\mu}_i$ are the scalar products of the observed direction $\underline{r}$ with three mutually orthogonal directions. (See Directional Distributions.)

Assuming that our data is fitted by one of these distributions, we may find maximum likelihood estimates and make likelihood ratio tests of various hypotheses in the usual way. If for the density exp K cos θ, the data is not too dispersed, it will be shown in §3 that those methods can be reduced to analogues of the familiar analysis of variance. These tests are appropriate when the K's are large and due to Watson (1956a,b, 1965). If the sample size is large, there are the usual simplified methods. For the general Beran family of densities, it is hard to estimate the parameters except with large samples.

It could well be that there is no preferred direction--for example, the pigeons may be unable to use any navigational clues and leave in random directions. A test for the stability of magnetization of rocks that left a formation (which would be magnetized uniformly) to be part of a conglomerate is that the direction of magnetization of pebbles in the latter is uniform. Thus, tests here for uniformity are perhaps of more practical importance than on the line.

The book by Mardia (1972) provides references to all the pre-1971 original papers and tables of significance points. Extensive references to Earth Science applications are given to Watson (1970). Kendall (1974), his references and the Discussion open up other related areas, practical and theoretical.

2. EXPLORATORY ANALYSIS

Data in two dimensions could be grouped by angle and sectors drawn with radii proportional to frequency. This "rose diagram" is the analogue of the histogram. One might use $(\text{frequency})^{\frac{1}{2}}$. In three dimensions, it is hard to view points on a sphere, so that projections are used. The equal area or Lambert projection makes the density of points easy to interpret. One may only see one hemisphere, so one tries to position it conveniently. Indeed, the ability to rotate the data freely and view projections is invaluable in practice. The programs are easy to write. Such plots will reveal the general shape of the data -- one or more clusters, points clustered around great circles, etc. On the sphere, 'histograms' are rarely used, but contouring methods are often used, e.g. one might compute a density estimate at $\underline{r}$ from

$$\hat{f}(\underline{r}) = \frac{1}{N} \sum_{i=1}^{N} w_N(\underline{r}'\underline{r}_i)$$

where $w_N(z)$ is a probability density on $(-1,1)$ suitably peaked at $z = 1$. As an example, we could use $w(r'r_i)$ proportional to $\exp K_N \underline{r}'\underline{r}_i$ where K_N increases with N.

The position of a single cluster is clearly suggested by $\hat{\underline{\mu}} = \underline{R}/R$ where $\underline{R} = \sum \underline{r}_i$ and R = length of $\underline{R}$. If the cluster is very dispersed, R will be much smaller than N so that $N - R$ is a measure of total dispersion of the sample. This suggests $(N-R)/N$ should be an analogue of the sample variance or dispersion. For example, if all the vectors are identical $R = N$, and this quantity is zero.

For axial data, data with diametrically opposed modes or clusters, and data around a great circle, one might look at

$$\sum \cos^2 \theta_i = \sum (\underline{r}_i' \underline{\nu})^2$$

as $\underline{\nu}$ varies over the sphere. Its stationary values are the eigen values of the matrix $\sum \underline{r}_i \underline{r}_i'$ and the eigen vectors interesting directions of $\underline{\nu}$. One large and two small eigen values suggest a single cluster or an axial distribution since there is a $\underline{\nu}$ that is nearly parallel or antiparallel to all the observations. A single small and two nearly equal large roots suggest a uniform distribution

equal. They are 195, 167, 144. But for Fig. 3 they are 119, 20, 13. The sums of these numbers, but for rounding errors, are 505 and 153 respectively. The vector resultant $\underline{R}$ of the vectors in Fig. 3 is (5·13, 1·36, 99·23) so its length $(5\cdot13^2 + 1\cdot36^2 + 99\cdot23^2)^{\frac{1}{2}} = 99\cdot37 = R$. The point where R meets the hemisphere is an estimate of the mean direction of these normals. $(153 - 99\cdot37)/153 = \cdot35$ is a measure of the dispersion of the data about this mean direction. The direction cosines of the mean direction are $(5\cdot13, 1\cdot36, 99\cdot23)/99\cdot37 = (0\cdot052, 0\cdot014, 0\cdot998)$. The eigen vector associated with the eigen value 119 is very similar (·071, ·018, ·997).

3. PARAMETRIC ANALYSES

For a single cluster on the sphere, it is reasonable to assume the Fisher distribution which yields a likelihood

$$\prod_{i=1}^{N} \frac{K}{4\pi \sinh K} \exp K\underline{r}_i'\underline{\mu} \quad \alpha \left(\frac{K}{\sinh \kappa}\right)^N \exp K\underline{R}'\underline{\mu} \ .$$

Thus the m.l. estimates are $\hat{\mu} = \underline{R}/R$ and $\hat{K}$ such that $\coth \hat{K} - 1/\hat{K} = R/N$ or approximately $k = (N-1)/(N-R)$. Thus for the data in Fig. 3 which seems (this could be examined more carefully) to follow the Fisher distribution, $\hat{\mu} = (0\cdot052, 0\cdot014, 0\cdot998)$ and $k = (153 - 1)/(153 - 99\cdot37)$

around a great circle whose normal is the eigen vector for the small root, etc. .

There is no severe problem here with "wild" observations, although they may affect measures of dispersion.

To illustrate the above suggestions, we consider data on the orbits of comets given in Marsden (1979). The orientation of the orbital planes and directions of motion may give clues on their origin. The normal to the orbit plane in the direction suggested by the right hand rule (fingers in the direction of motion, thumb indicating the normal) is a unit vector. Looking down onto the plane of the elliptic, and using an equal area projection, the vectors associated with all periodic comets are shown in Fig. 1. The clumping of 658 points in the center (or pole of the hemisphere) indicates that many comets move like the planets in orbits near the plane of the earth's orbit. Cometary orbits change so we have used orbits associated with their last apparition, or sighting. If only the 505 comets with periods greater than 1000 years are plotted (see Fig. 2), their normals appear to be uniformly distributed. The distribution of the 153 normals to the orbital planes of comets with periods of less than 1000 years (see Fig. 3) is concentrated. The superposition of Figs. 2 and 3 is of course Fig. 1. The eigen values of the matrix $\sum r_i r_i'$ for the data in Fig. 2 are, as should be expected, fairly

= 2·87 . These are the intuitive estimators derived above. Writing $\underline{r}_i'\underline{\mu} = \cos\theta_i$, and letting K be large so $2 \sinh K \doteq \exp K$, one may show that $2K(1 - \cos\theta)$ is distributed like χ^2_2 . Hence, if $\underline{\mu}$ is known

$$\sum 2K\,(1-\cos\theta_i) = 2K(N-R'\underline{\mu}) \quad \textit{is distributed like} \quad \chi^2_{2N} .$$

One might guess that when $\underline{\mu}$ is fitted to the data, 2 d.f. will be lost so that $2K(N-R)$ is approximately $\chi^2_{2(N-1)}$. Hence, we may write, setting $R'\mu = X$, the identity

$$N-X = N-R + R-X$$

and give it the interpretation

dispersion about true $\underline{\mu}$ = *dispersion about estimate* $\hat{\underline{\mu}}$ + *dispersion of* $\hat{\underline{\mu}}$ *about* $\underline{\mu}$.

Continuing this analogue to the analysis variance, we have:

$$2K(N-X) = 2K(N-R) + 2K(R-X) ,$$
$$\chi^2_{2N} = \chi^2_{2(N-1)} + \chi^2_2 ,$$

so that the test of a prescribed mean $\underline{\mu}$ is provided by

$$\frac{\textit{dispersion of } \hat{\underline{\mu}} \textit{ about } \underline{\mu}}{\textit{dispersion about } \hat{\underline{\mu}}} = \frac{2K(R-X)}{2K(N-R)} = \frac{\chi^2_2}{\chi^2_{2(N-1)}} .$$

Thus one may use $F_{2,2(N-1)}$ to make the test. The reader should examine Fig. 4 to see the commonsense of the text.

The data in Fig. 3 have a k of only 2·87, about the minimum for which the above approximation makes sense. To test the null hypothesis that the true mean normal is perpendicular to the ecliptic, i.e. direction $\underline{\mu} = (0,0,1)$, $\underline{R}'\underline{\mu} = 99{\cdot}23 = X$. Thus $(N-1)(R-X)/(N-R) = 0{\cdot}04$ is very small compared to $F_{2,306}$ -- clearly the null hypothesis is strongly supported.

Similarly, to test that two populations (with the same large K) have the same mean direction given samples sizes N_1, N_2 and resultants $\underline{R}_1$ and $\underline{R}_2$, Fig. 5 suggests the statistic

$$\frac{R_1+R_2-R}{(N_1-R_1) + (N_1-R_1)}$$

and the identities (with $N = N_1+N_2$, $\underline{R} = \underline{R}_1+\underline{R}_2$)

$$2K(N-R) = 2K(N_1-R_1) + 2K(N_2-R_2) + 2K(R_1+R_2-R) ,$$

$$\chi^2_{2(N-1)} = \chi^2_{2(N_1-1)} + \chi^2_{2(N_2-1)} + \chi^2_2 = \chi^2_{2(N-2)} + \chi^2_2$$

suggest that $F_{2,2(N-2)}$ may be used.

Similar tests (with tables) are available on the circle for the von Mises. Details of these, some exact and further approximate tests for all the distributions above are given in Mardia's book. Much of this work is due to M.A. Stephens. The result is a fairly complete set of analogues of normal tests for independent observations. Conspicuously lacking so far are satisfactory analogues for correlated directions (but see Stephens [1979]) and time series or spatial fields of directions. Fortunately, such problems seem rare in practice. Wellner (1978) extends the two sample theory -- see his references to related work. Bingham (1974) gives methods for his distribution.

4. TESTS OF UNIFORMITY

As mentioned earlier, with reference to the homing directions of disoriented pidgeons, the direction of magnetization of pebbles in a conglomerate and the normals to the orbits of comets with periods over 1000 years (see Fig. 2), the problem of testing for uniformity (q.v.) arises more often here than on the line. One has an intuitive feeling whether a set of points on a circle, or sphere, or an equal

area projection of a sphere suggest non-uniformity in the population they come from, and intuition suggests test statistics.

If the population is unimodal, the length R of the vector resultant should be longer than it would be when sampling from a uniform parent. The Fisher and von-Mises unimod distributions become uniform when K=0. R does not depend upon the coordinate system--it is invariant. Hypothesis testing (a.v.) theory shows that in this case, the best invariant test of K=0 is Rayleigh's test: reject uniformity if R is significantly large. Now

$$R^2 = (\Sigma x_i)^2 + (\Sigma y_i)^2 + (\Sigma z_i)^2$$

where $(x_i, y_i, z_i) = \underline{r}_i$. When the $\underline{r}_i$ are independently uniformly distributed over the sphere

$$Ex_i = Ey_i = Ez_i = 0 \quad ,$$
$$E(x_i y_i) = E(y_i z_i) = E(z_i x_i) = 0 \quad ,$$
$$E(x_i^2) = E(y_i^2) = E(z_i^2) = \frac{1}{3}$$

Then Σx_i, Σy_i, Σz_i become independently Gaussian, means zero, variances N/3 and R^2 becomes, on the null hypothesis, $N\chi_3^2/3$. On the circle $R^2 = (\Sigma x_i)^2 + (\Sigma y_i)^2$ is, by a similar argument, asymptotically $N\chi_2^2/2$. These are also likelihood ratio tests.

If the distribution is antipodally symmetric, R is clearly not powerful. The likelihood ratio test for a

Bingham alternative might then be used--see Bingham (1974). When this test is used on the data in Fig. 2, it shows that the points are significantly non-random! The eye is a poor judge in the other direction too--one often thinks one sees features in purely random data.

One may ask for tests that are in the spirit of the Kolmogorov and Cramér-von Mises tests. While one may choose a starting point on the circle and form the sample distribution, the resulting tests depend upon the starting point, i.e., they are not invariant. Invariant tests for the circle were first constructed intuitively; Kuiper (1960) gave an invariant form of the Kolmogoroff-Smirnov tests and Watson (1961, 1962) gave U^2, an invariant form of the Cramér-von Mises tests.

Motivated by Ajne (1968), Beran (1968) discovered a very general theory to derive statistics of the U^2 type on homogeneous spaces as locally optimal tests. See also Giné (1975) and Prentice (1978) for further generalizations. Wellner's (1978, 1979) 2 sample (asymptotic) tests arise by applying the permutation idea to the Fourier co-efficients used since Watson (1961), Beran (1968) in this literature on uniformity testing. (See also Goodness-of-Fit Tests.) Watson (1974) produced Kuiper-Kolmogoroff type tests as optimal tests for distant, not local, alternatives.

5. MISCELLANEOUS REMARKS

These topics flow naturally into more general orientation problems--we have dealt with the orientation of a line or arrow, but a solid body is oriented by an orthogonal matrix. They also raise particular cases of the fascinating problem of finding suitable definitions on new manifolds of familiar quantities like means, dispersions, and correlations.

More references to modern work can be traced through Wellner (1979), and Beran (1979) who exploits the exponential family simplification; Beran avoids the complex maximum likelihood estimation by using a non-parametric estimator of the logarithm of the density.

The probability and statistics of directed quantities has some very early history. Buffon solved his needle problem in 1733. Daniel Bernoulli tried in 1734 to show that it is very unlikely that the new coincidence of the planetary orbits is an accident. (See, e.g., Watson, 1970, 1977).

Acknowledgement: We wish to thank Elizabeth Ryder for analyzing Marsden's comet data.

REFERENCES:

AJNE, B. (1968). A simple test for uniformity of a circular distribution. Biometrika 55, 343-354.

BERAN, R.J. (1968). Testing for uniformity on a compact homogeneous space. J. Appl. Prob. 5, 177-95.

BERAN, R.J. (1969). The derivation of two-sample tests from tests for uniformity of a circular distribution

BERAN, R.J. (1974). Exponential models for directional data. Ann. Math. Stat. 7, 1162-1179.

BINGHAM, C. (1974). An antipodally symmetric distribution on the sphere. Ann Math Stat. 6, 292-313.

FISHER, R.A. (1953). Dispersion on a Sphere. Proc. Roy. Soc. Lond. A217, 195-305. 220, 230, 240, 242, 244, 251, 261, 283

KENDALL, D.G. (1974). Pole-seeking Brownian motion and Bird Navigation. J.R.S.S. B, 36, 365-417.

KRUMBEIN, W.C. (1939). Preferred orientation of pebbles in sedimentary deposits. J. Geol. 47, 673-706. [10, 70]

KUIPER, N.H. (1960). Tests concerning random points on a circle. Nederl. Akad. Wetensch. Proc. Ser A 63, 38-47.

MARDIA, K.V. (1972). Statistics of Directional Data, Academic Press.

MARSDEN, B.G. (1979). Catalogue of Cometary Orbits. Cambridge, Mass: Smithsonian Astrophysical Observatory.

PRENTICE, M.J. (1978). On invariant tests of uniformity for directions and orientations. Ann. Statist. 6, 169-176.

STEPHENS, M.A. (1979). Vector Correlation. Biometrika 66, 41-48.

WATSON, G.S. (1956a). Analysis of dispersion on a sphere. Monthly Notices Roy. Astro. Soc., Geophys. Supp. 7, 153-159.

WATSON, G.S. (1956b). A test for randomness of directions. Monthly Notices Roy. Astr. Soc., Geophys. Supp. 7, 160-161.

WATSON, G.S. (1961). Goodness-of-fit tests on a circle, Biometrika 48, 109-114.

WATSON, G.S. (1962). Goodness-of-fit tests on a circle, II, Biometrika 49, 57-63.

WATSON, G.S. (1965). Equatorial distributions on a sphere. Biometrika 52, 193-201.

WATSON, G.S. (1967a). Another test for the uniformity of a circular distribution. Biometrika 54, 675-677.

WATSON, G.S. (1970). Orientation statistics in the earth sciences. Bull. Geol. Inst., Univ. Uppsala 2, 73-89.

WATSON, G.S. (1974). Optimal invariant tests for uniformity. "Studies in Probability and Statistics", Jerusalem Acad. Press, 121-128.

WELLNER, JOHN A. (1978). Two-sample tests for a class of distributions on the sphere. Unpublished manuscript.

WELLNER, JOHN A. (1979). Permutation Tests for Directional Data. The Ann. of Stat. 7, 5, 929-943.

CAPTIONS

Fig. 1. Equal area plot of the normals to the last seen orbit of 658 comets, as seen looking vertically down onto the ecliptic

Fig. 2. Equal area plot of the normals to the last seen orbit of all 505 comets with periods greater than 1000 years, as seen looking vertically down onto the ecliptic

Fig. 3. Equal area plot of the normals to the last seen orbit of the 153 comets with periods less than 1000 years, as seen looking vertically down onto the ecliptic

Fig. 4. The sample vector resultant $\underline{R}$ whose projection down onto the hypothetical mean direction $\underline{\mu}$ has length X. If $\underline{R}$ and $\underline{\mu}$ are nearly parallel (orthogonal), R-X will be small (large)

Fig. 5. The sum of the lengths vector resultants $\underline{R}_1$ and $\underline{R}_2$ of two samples will be only slightly (much) larger than the length of $\underline{R}$, the resultant of all the data if $\underline{R}_1$ and $\underline{R}_2$ are nearly parallel (greatly inclined)

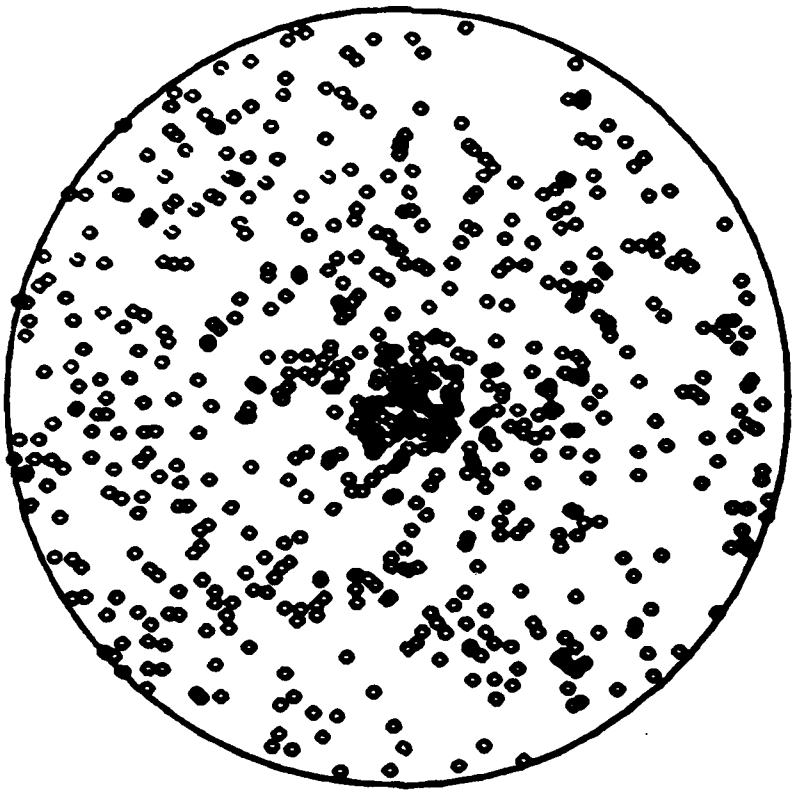


Fig. 1

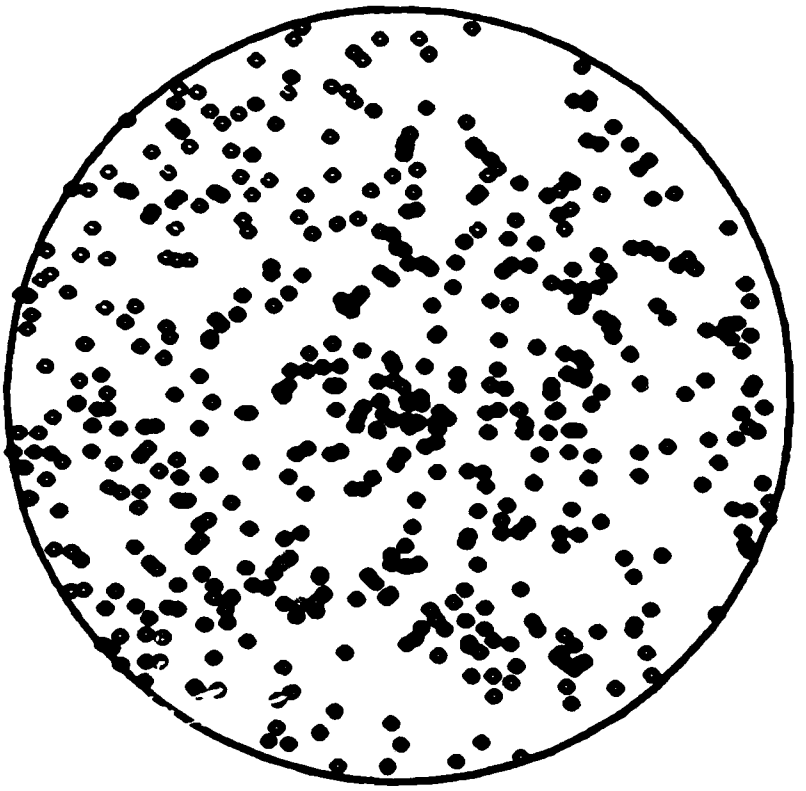


Fig. 2

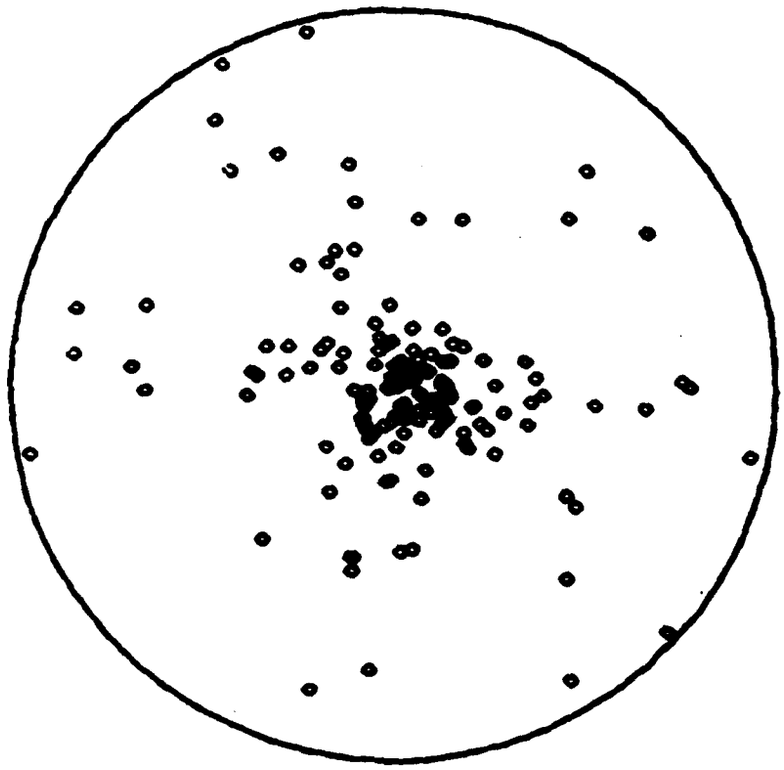


Fig. 3

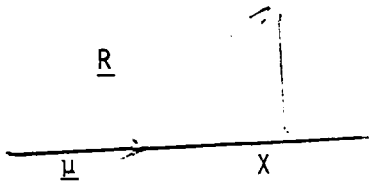


Fig. 4

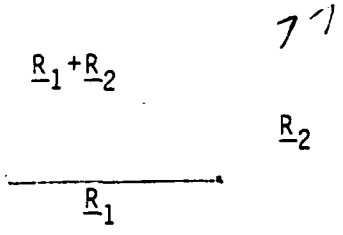


Fig. 5